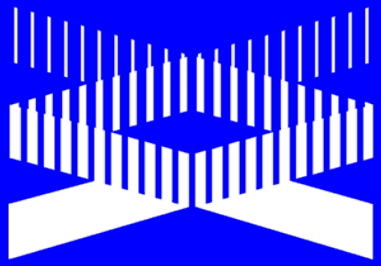
Epidemiology for Researchers Performing Genetic/Genomic Studies:

Application of Epidemiologic Methods to Human Genome Research



Essay

Why Most Published Research Findings Are False

John P. A. Ioannidis

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when

factors that influence this problem and some corollaries thereof.

Modeling the Framework for False Positive Findings

Several methodologists have pointed out [9–11] that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a *p*-value less than 0.05. Research is not most appropriately represented and summarized by *p*-values, but, unfortunately, there is a widespread

is characteristic of the field and can vary a lot depending on whether the field targets highly likely relationships or searches for only one or a few true relationships among thousands and millions of hypotheses that may be postulated. Let us also consider, for computational simplicity, circumscribed fields where either there is only one true relationship (among many that can be hypothesized) or the power is similar to find any of the several existing true relationships. The pre-study probability of a relationship being true is $R/(R + 1)$. The probability of a study finding a true relationship reflects the power $1 - \beta$ (one minus the Type II error rate). The probability

PLoS Med. 2005 Aug;2(8):e124.

WSJ.com THE WALL STREET JOURNAL. ONLINE

SCIENCE JOURNAL

By ROBERT LEE HOTZ

Most Science Studies Appear to Be Tainted By Sloppy Analysis

September 14, 2007; Page B1

We all make mistakes and, if you believe medical scholar John Ioannidis, scientists make more than their fair share. By his calculations, most published research findings are wrong.

Dr. Ioannidis is an epidemiologist who studies research methods at the University of Ioannina School of Medicine in Greece and Tufts University in Medford, Mass. In a series of influential analytical reports, he has documented how, in thousands of peer-reviewed research papers published every year, there may be so much less than meets the eye.

WSJ. 2007Sep14.

Course Purpose and Goals

- Purpose: Description of the theory and methods of epidemiology that are applicable to human genome research
- Goals
  - Optimal application of modern genome analysis methodologies to studies of unrelated subjects in human populations
  - Use of epidemiologic studies most appropriate to answer the genomic question
    - Study design
    - Collection of data
    - Interpretation of results

Course Faculty

Co-Directors:

Teri Manolio, MD, PhD (Office of Population Genomics, NHGRI)
Thomas Pearson, MD, PhD (Univ. of Rochester CTSI)

Faculty:

Emily Harris, PhD (Office of Population Genomics, NHGRI)
Lucia Hindorff, PhD (Office of Population Genomics, NHGRI)
Erin Ramos, PhD (Office of Population Genomics, NHGRI)
Jeffery Struewing, MD (Office of Population Genomics, NHGRI)

Organizers:

Mia Diggs (Office of Population Genomics, NHGRI)
Lisa McNeil (Office of Population Genomics, NHGRI)

CONFLICT OF INTEREST DISCLOSURE

Larson, G. *The Complete Far Side*. 2003.

Overview of Course

- Learning objectives
- Format
  - Eight lectures
  - Applications
  - Discussion
  - Webcast

Course Outline

1. Course Overview	Thomas Pearson
2. Measuring Phenotypes	Erin Ramos
3. Measures of Association and Risk	Emily Harris
Break	
4. Epidemiologic Study Designs	Lucia Hindorff
Lunch	

Course Outline (Cont'd)

Lunch	
5. Study Replication	Teri Manolio
6. Bias in Human Genome Research	Teri Manolio
Break	
7. Genetic Screening and Diagnosis	Jeffery Struewing
8. Practical Applications	Thomas Pearson
Wrap-up/Course Evaluation	T. Manolio/T. Pearson

NHGRI Catalog of GWAS (www.genome.gov/gwastudies/)

- All publications reporting genome-wide association studies (beginning March, 2005)
  - Platforms with density of at least 100,000 SNPs
  - Identified by literature searches, media, HUGE Navigator
- Data Presented
  - Citation
  - Disease/trait
  - Sample sizes
  - Chromosomal region
  - Gene
  - Associations
    - Significant risk alleles
    - Odds ratio per copy
    - Risk allele frequency
    - P value of association

Lecture 1: Epidemiology for Geneticists Versus Genetic Epidemiology

Thomas A. Pearson, MD, PhD
University of Rochester School of Medicine
Visiting Scientist, NHGRI (9/1/07-5/30/08)

Lecture 1: Learning Objectives

- Provide an overview of the uses of epidemiology in genomic research
- Review current methods for measuring genetic exposures associated with common, complex diseases
- Review current methods for measuring environmental exposures associated with common, complex diseases
- Emphasize the population perspective, rather than the individual subject's perspective

Introduction to Epidemiology

- Definition: The study of how disease is distributed in populations and the factors that influence or determine this distribution
- Key Assumption: Disease is not randomly distributed throughout the population
- The Epidemiologic Method:
  - Determine if an association between a factor or a characteristic exists with a disease
  - Derive inferences regarding a possible causal relationship from patterns of associations found

Frequently Asked Questions

By Patients:

- What is the disease and how common is it?
- What are my chances of a bad outcome?
- What caused this disease?
- Is there a treatment and will it help me?

By Policy Makers (in addition to #'s 1-4):

5. What are the implications of the disease to clinical care and public health programs?

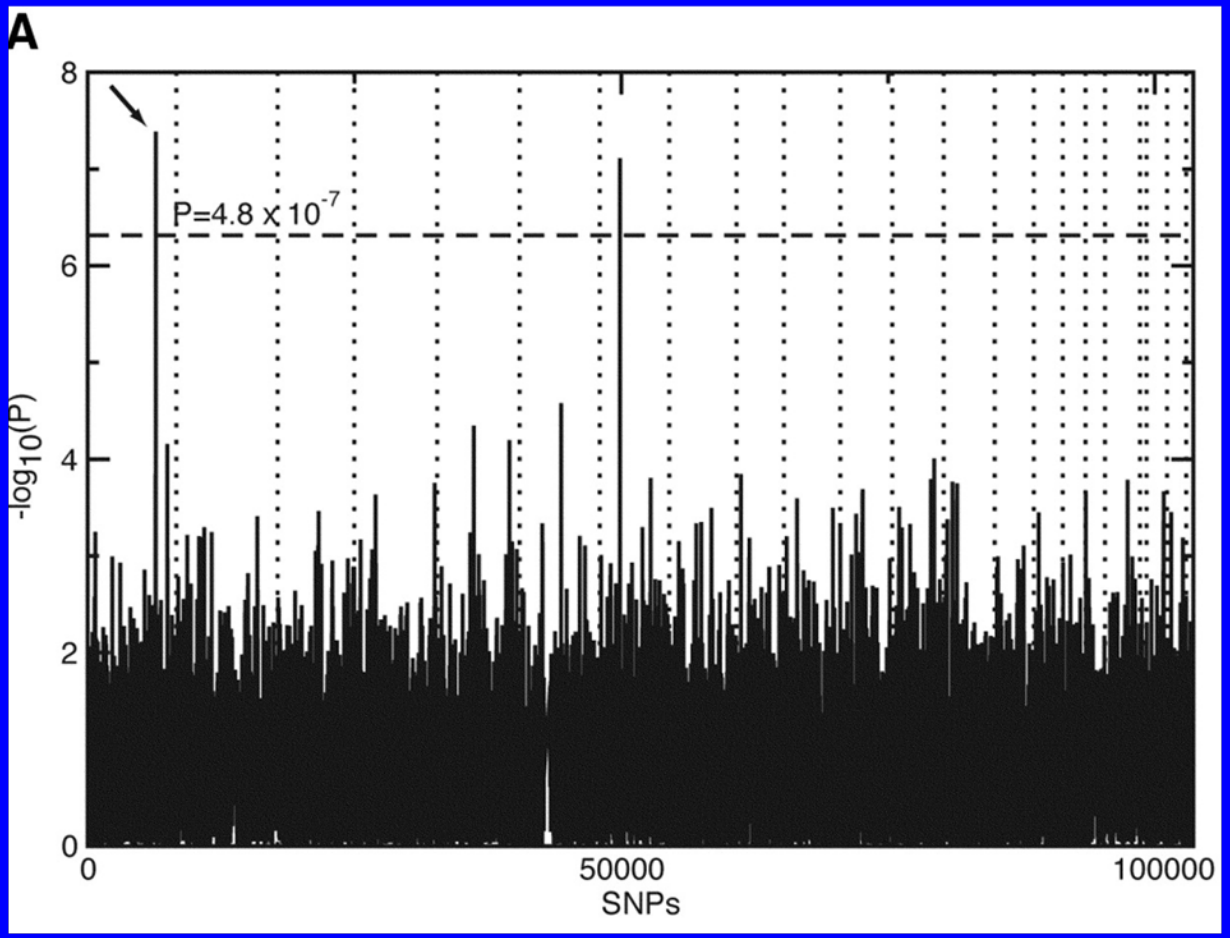
Objectives of Epidemiology

1. To determine the extent of disease found in the community
2. To study the natural history and prognosis of the disease or condition
3. To identify the etiology or the cause of the disease – its risk factors – that is, factors that increase a person's risk for disease
4. To evaluate new preventive and therapeutic measures and new modes of healthcare delivery
5. To provide the foundation for developing public policy and regulatory decisions related to exposures

Genomic Analysis and Association Studies

- 1900-present: Human genetics, Mendelian disorders.
- 1953-present: Molecular genetics, structure and function of the human genome.
- 1980-present: Identification of genetic variants and candidate gene studies.
- 2003-present: Sequencing of the entire human genome.
- 2005-present: Genome-wide association studies.

P Values of GWA Scan for Age-Related Macular Degeneration



Klein et al, *Science* 2005; 308:385-389.

Odds Ratios and Population Attributable Risks for AMD

Attribute (SNP)	rs380390 (C/G)	rs1329428 (C/T)
Risk allele	C	C
Allelic association χ^2 P value	4.1×10^{-8}	1.4×10^{-6}
Odds ratio (dominant)	4.6 [2.0-11]	4.7 [1.0-22]
Frequency in HapMap CEU	0.70	0.82
Population Attributable Risk	70% [42-84%]	80% [0-96%]
Odds ratio (recessive)	7.4 [2.9-19]	6.2 [2.9-13]
Frequency in HapMap CEU	0.23	0.41
Population Attributable Risk	46% [31-57%]	61% [43-73%]

Klein et al, *Science* 2005; 308:385-389.

Familial Hypercholesterolemia

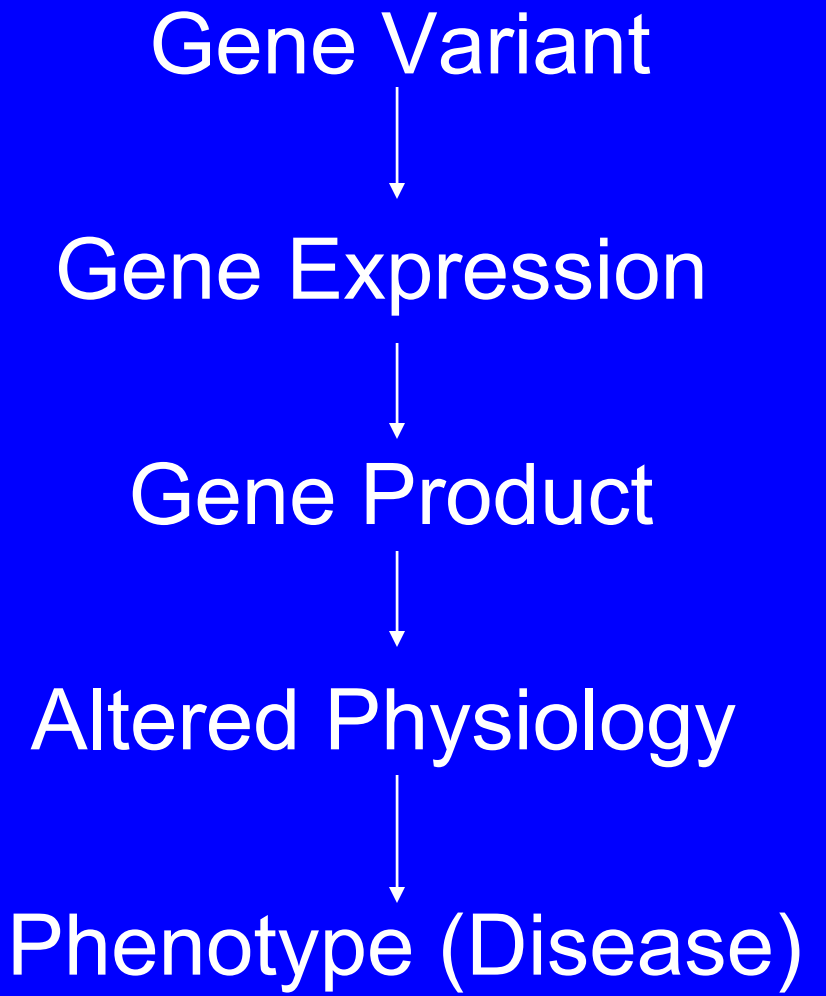
- Caused by polymorphisms affecting the low density lipoprotein receptor
- Autosomal dominant inheritance pattern
  - Heterozygous: elevated serum LDL cholesterol (>300 mg/dl); vascular disease in middle age
  - Homozygous: extreme LDL cholesterol elevations (>700 mg/dl); vascular disease in childhood
- Prevalence of FH alleles
  - U.S.: 1:500
  - French Canadian: 1:82
  - S. Africa Afrikaners: 1:62

BRCA1 and BRCA2: Estimated Lifetime Risk of Cancer*

	Breast	Ovarian
BRCA1	65% (44-78%)	39% (18-54%)
BRCA2	45% (31-56%)	11% (2-19%)

Antoniou, et al. ASHG 2003; 72: 1117

The Genetic Etiology of Disease

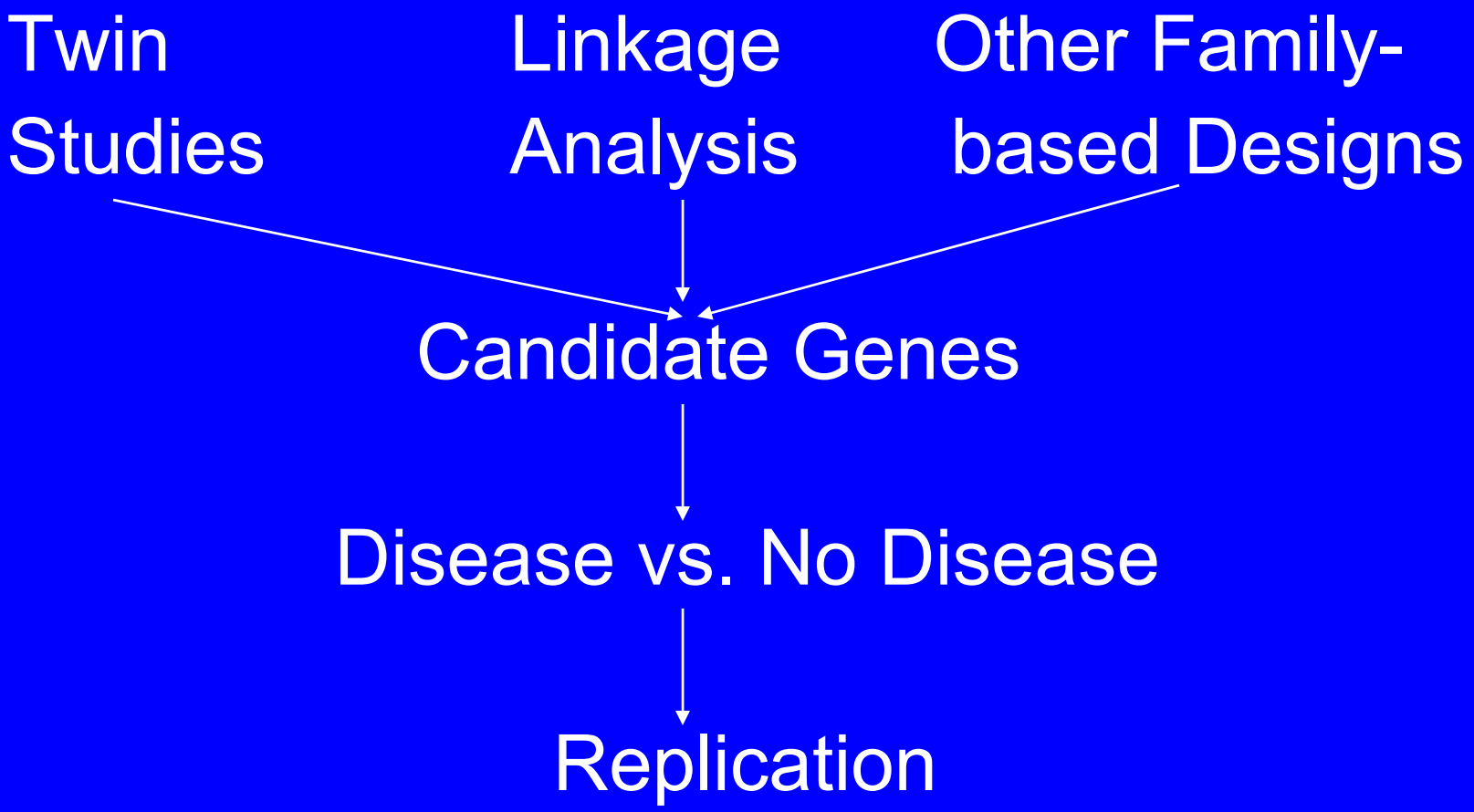


Patterns of Inheritance

- Mendelian Disease: Condition (phenotype) caused almost entirely by a single major gene, in which the disease is manifested in only 1 (recessive) or 2 (dominant) of the 3 possible genotype groups.
- Common disease, Common Variant:

  Common conditions (phenotype) attributable to a limited number of allelic variants which occur in 1-5% or more of the population.

Candidate Gene Approaches (Hypothesis-driven)



Genome-wide Association (Agnostic)

Entire Genome

↓

Disease vs. No Disease

↓

Replication

Susceptibility Variants Associated with Systemic Lupus Erythematosus in Women*

- Case-control study of 720 women with SLE and 2337 control women.
- 317,501 SNPs assessed genome-wide
- Two replication studies with 1846 female cases and 1825 female controls.
- At least 17 SNPs associated with SLE at P<2xE-7

*International Consortium for SLE Genetics. Nature Genetics 1/20/08

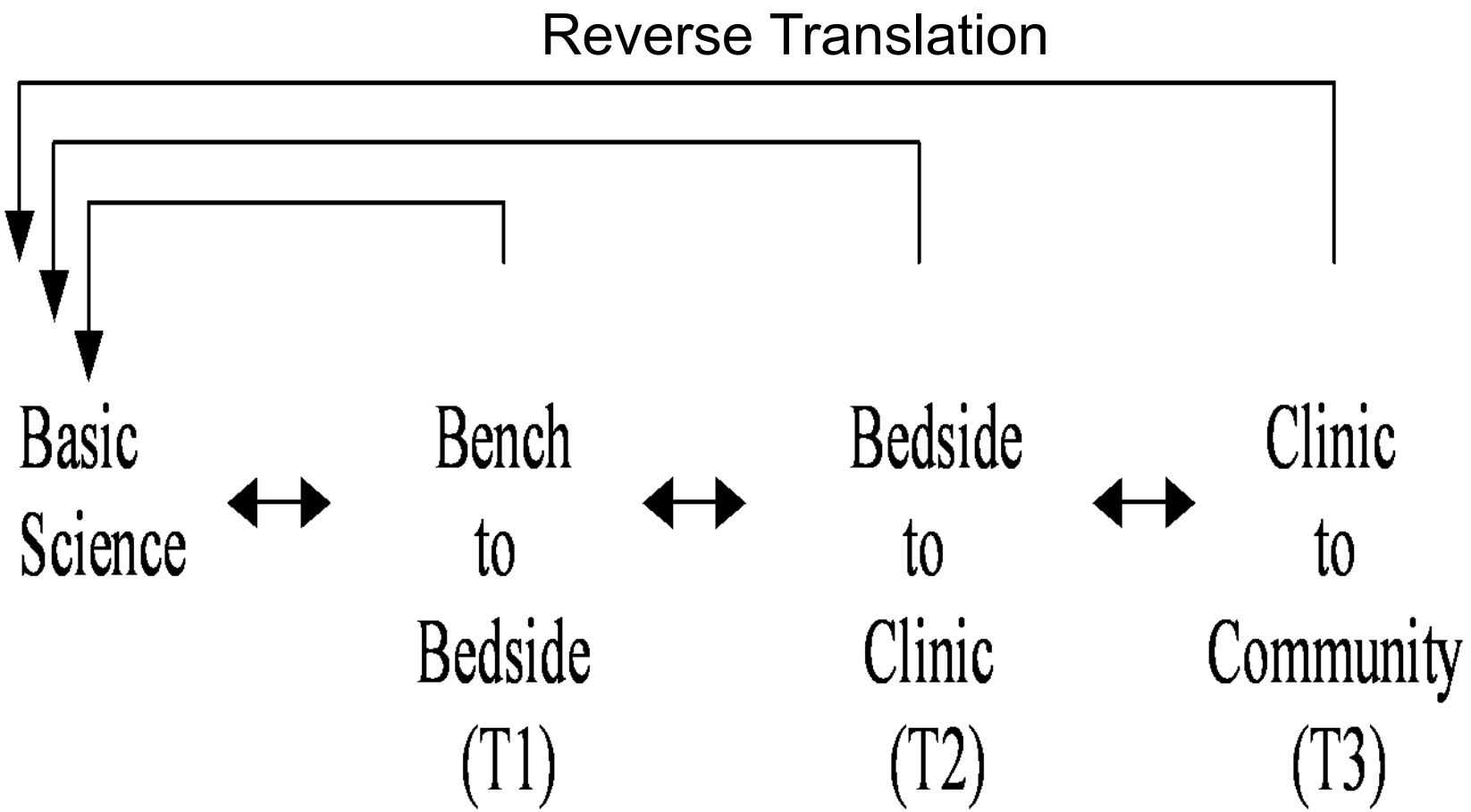
Logistic Regression Model of Independent Contributions of Markers Associated with SLE*

Gene	Chromosome	OR	P
PKY	3p14.3	1.27	9.2E-07
HLA region	6p21.33	1.82	4.5E-17
HLA region	6p21.32	1.40	2.8E-12
IRF5/TNP03	7q32.1	1.61	1.7E-14
KIAA1542	11p15.5	0.78	1.3E-07
ITGAM	16p11.2	1.70	1.9E-18

Cstatistic=0.67;15% of heritability explained

*Int. Consort. for SLE Genetics.NatGen 1/20/08

Translational Research
Reverse Translation



Lessons Learned from Initial GWA Studies

Signals in Previously Unsuspected Genes

Macular Degeneration	*CFH*
Coronary Disease	*CDKN2A/2B*
Childhood Asthma	*ORMDL3*
Type II Diabetes	*CDKAL1*
QT interval prolongation	*NOS1AP*

Lessons Learned from Initial GWA Studies

Signals in Previously Unsuspected Genes	
Macular Degeneration	*CFH*
Coronary Disease	*CDKN2A/2B*
Childhood Asthma	*ORMDL3*
Type II Diabetes	*CDKAL1*
QT interval prolongation	*NOS1AP*
Signals in Gene "Deserts"	
Prostate Cancer	8q24
Crohn Disease	5p13.1, 1q31.2, 10p21

Lessons from GWA Studies

Signals in Previously Unsuspected Genes

Macular Degeneration	*CFH*
Coronary Disease	*CDKN2A/2B*
Childhood Asthma	*ORMDL3*
Type II Diabetes	*CDKAL1*
QT Interval Prolongation	*NOS1AP*

Signals in Gene "Deserts"

Prostate Cancer	8q24
Crohn Disease	5p13.1, 1q31.2, 10p2

Signals in Common

Diabetes, CHD, Melanoma	*CDKN2A/2B*
Prostate, Breast, CR Cancers	8q24 region
Crohn's Disease, Psoriasis	*IL23R*

Genome-wide Association and Clinical Trials

Genome-wide pharmacogenetic investigation of a hepatic adverse event without clinical signs of immunopathology suggests an underlying immune pathogenesis

5/2007

A Kindmark[1], A Jawaid[2], CG Harbron[2], BJ Barratt[2], OF Bengtsson[1], TB Andersson[1], S Carlsson[1], KE Cederbrant[3],

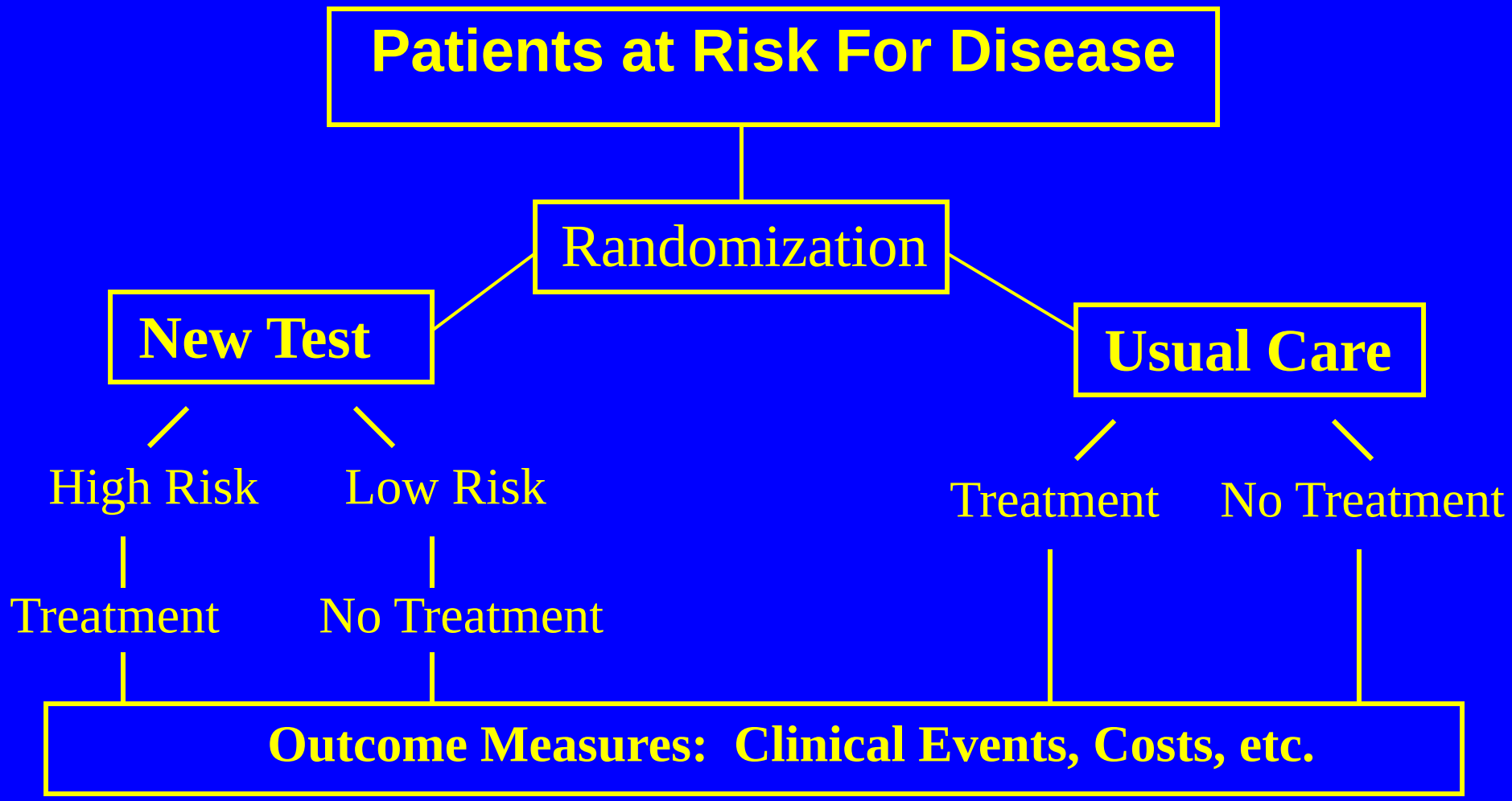
One of the major goals of pharmacogenetics is to elucidate mechanisms and identify patients at increased risk of adverse events (AEs). To date, however, there have been only a few successful examples of this type of approach. In this paper, we describe a retrospective case–control pharmacogenetic study of an AE of unknown mechanism, characterized by elevated levels of serum

Genome-Wide Pharmacogenomic Analysis of the Response to Interferon Beta Therapy in Multiple Sclerosis

1/2008

Esther Byun, MD; Stacy J. Caillier, BSc; Xavier Montalban, MD; Pablo Villoslada, MD, PhD; Oscar Fernández, MD; David Brassat, MD; Manuel Comabella, MD, PhD; Joanne Wang, MPH; Lisa F. Barcellos, PhD; Sergio E. Baranzini, PhD; Jorge R. Oksenberg, PhD

Cost – Effectiveness Trial of New Biomarker Test



Personalized Medicine

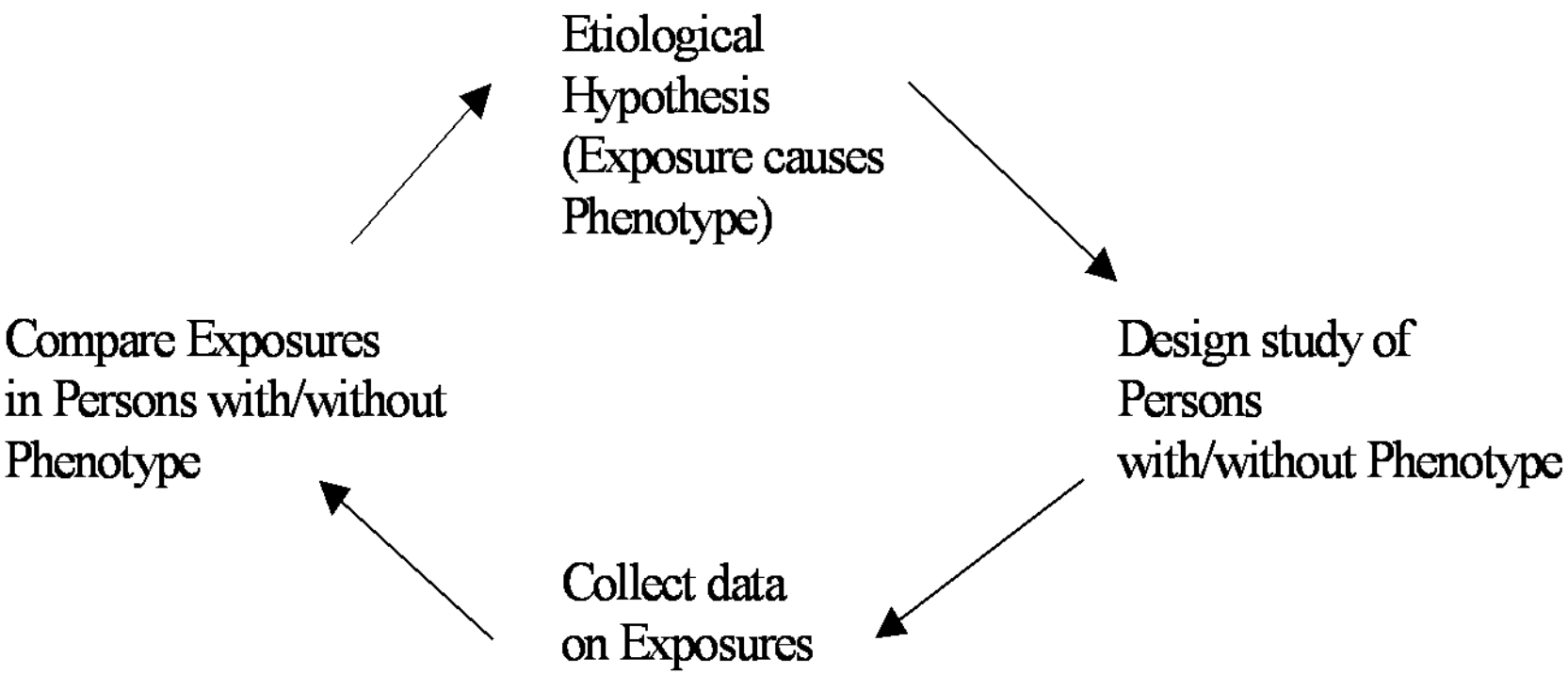
"At its most basic, personalized medicine refers to using information about a person's genetic make-up to tailor strategies for detection, treatment, and prevention of disease"

Francis Collins,
Director, NHGRI
7/17/05

GINA: The Genetic Information Non-Discrimination Act 2007-2008

- Prohibits health insurers from requesting or requiring genetic information of an individual or their family members or using it for decisions on coverage, rates, etc.
  - Includes participation in research that includes genetic services
- Prohibits employers from requesting or requiring information or using it in decisions regarding hiring, firing, or terms of employment

The Epidemiologic Method



Overview of Epidemiologic Design Strategies

- Descriptive studies
  - Populations (correlational studies)
  - Individuals
    - Case reports
    - Case series
    - Cross-sectional surveys
- Analytic studies
  - Observational studies
    - Case-control studies
    - Cohort studies-retrospective and prospective
  - Intervention studies (clinical trials)

Measuring Genetic Exposures

Restriction-fragment length polymorphisms

Variable number of tandem repeats

Single nucleotide polymorphisms

Sequencing

Gene expression

Gene products (e.g. blood groups)

Epigenetics

Quality Control of SNP Genotyping: Samples

- Identity with forensic markers (Identifiler)
- Blind duplicates
- Gender checks
- Cryptic relatedness or unsuspected twinning
- Degradation/fragmentation
- Call rate (> 80-90%)
- Heterozygosity: outliers
- Plate/batch calling effects

Chanock et al, *Nature* 2007; Manolio et al *Nat Genet* 2007

Quality Control of SNP Genotyping: SNPs

- Duplicate concordance (CEPH samples)
- Mendelian errors (typically ≤ 1)
- Hardy-Weinberg errors (often > 10^{-5})
- Heterozygosity (outliers)
- Call rate (typically > 98%)
- Minor allele frequency (often > 1%)
- Validation of most critical results on independent genotyping platform

Chanock et al, *Nature* 2007; Manolio et al *Nat Genet* 2007

Coverage, Call Rates, and Concordance of Perlegen and Affymetrix Platforms on HapMap Phase II

Metric	Perlegen		Affymetrix	
No. of SNPs	480,744		439,249	
Coverage	Single Marker	Multi-Marker	Single Marker	Multi-Marker
CEU	0.90	0.96	0.78	0.87
CHB + JPT	0.87	0.93	0.78	0.86
YRI	0.64	0.78	0.63	0.75
Ave. call rate	98.9%		99.3%	
Concordance				
Homozygous genotypes	99.8%		99.9%	

GAIN Collaborative Group, Nat Genet 2007; 39: 1045-51

The Common Disease-Common Variant Hypothesis

SNP1 in Exon1 of GeneA

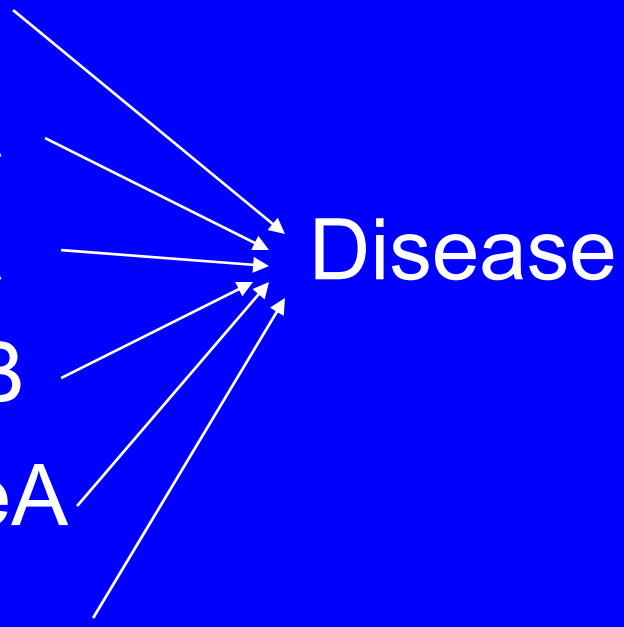
SNP1 in Exon 3 of GeneA

SNP2 in Exon 3 of GeneA

SNP 2 in Exon 2 of GeneB

SNP, Reg. Element, GeneA

Environmental Exposures



Nongenetic Exposures

Environmental Exposures (Air pollution, radiation)

Behaviors (Diet, Exercise, Tobacco)

Therapeutics (Drugs, Devices)

Risk of Developing AMD by CFH Y402H and Modifiable Risk Factors

Risk Factor	CFH Y402H Genotype		
	YY	YH	HH
BMI < 30 kg/m^2	1.00	1.95 [1.42-2.67]	3.96 [2.69-5.82]
BMI ≥ 30 kg/m^2	1.98 [0.91-4.31]	2.19 [1.11-4.30]	12.28 [4.88-30.90]
Non-smoker	1.00	1.95 [1.41-2.71]	4.23 [2.86-6.27]
Current smoker	2.34 [1.20-4.55]	3.20 [1.85-5.55]	8.69 [3.86-19.57]

Schaumberg DA et al, *Arch Ophthalmol* 2007; 125:55-62.

Challenges in Studying Gene-Environment Interactions

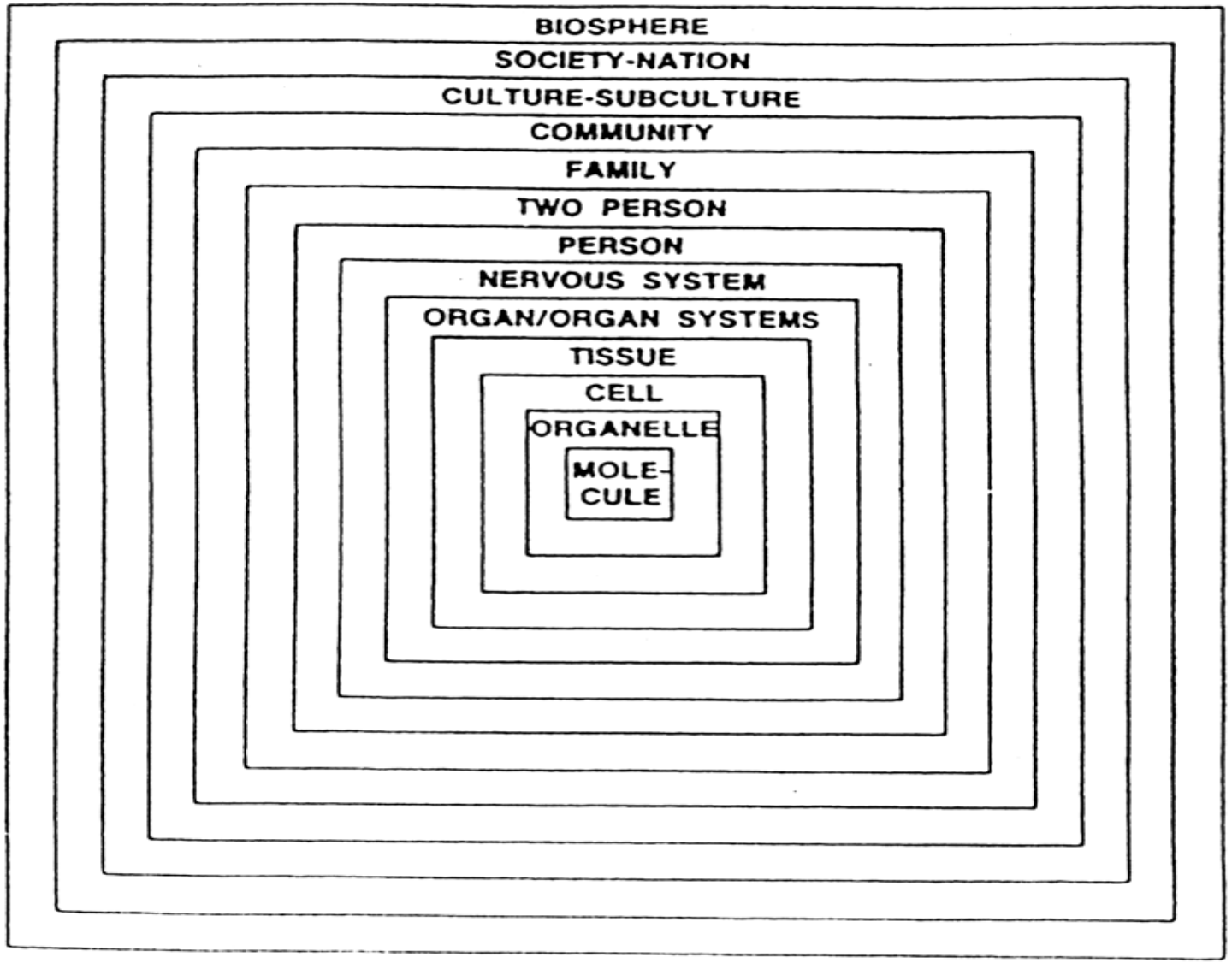
Challenge	Genes	Environment
Ease of measure	Pretty easy	Often hard
Variability over time	Low/none	High
Recall bias	None	Possible
Temporal relation to disease	Easy	Hard

Application

- Consider the gene/biological system on which the work of your laboratory focuses:
  - How common are genetic polymorphisms and related diseases?
  - Do organisms with the polymorphism have an altered natural history?
  - Is the polymorphism associated with phenotype or disease?
  - Is the gene or its products a target for intervention?
  - Should the polymorphisms be measured, and if so, why?

Biopsychosocial Model of George Engel, M.D.

Continuum of Natural Systems



Summary Points

- Epidemiology provides a population perspective important to the interpretation of genome-phenotype associations
- Epidemiologic methods are used to establish associations between possible exposures, including genomic variants, and disease
- Reverse translation has been a major contribution of genome-wide association studies
- Measurement of exposures, genomic or environmental, require rigorous quality control

Questions?